

AMENDMENTS TO THE CLAIMS

1 1. (Currently Amended) A method of categorizing a plurality of new electronic
2 documents into a set of categories, comprising the steps of:
3 establishing a plurality of training sets, wherein each training set is associated with a
4 category and includes training documents that have been classified as
5 belonging to said associated category;
6 determining how strongly each document of said plurality of documents corresponds
7 to each of said plurality of categories by determining similarity between said
8 each document and the training documents that belong to the training set of
9 said category; and
10 wherein the step of determining similarity is performed using a matrix representing
11 document similarity that is derived by combining two or more measures of
12 document similarity and by one calculation of a function referencing attributes
13 associated with each of said plurality of documents.

1 2. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity include hyperlink similarity.

1 3. (Original) A method as recited in Claim 2, in which two documents among the
2 plurality of documents are considered similar to each other when there is a
3 link from one to the other, or when the two documents link to, or are linked to
4 by, a set of other associated documents.

1 4. (Original) A method as recited in Claim 3, in which certain hyperlinks have
2 greater or lesser similarity weight than other hyperlinks, based on other
3 features of the links or their source or destination documents.

1 5. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity include a similarity of text of the documents.

1 6. (Original) A method as recited in Claim 5, wherein two documents are
2 considered similar based on a comparison of word vectors derived from the
3 text of each of the two documents.

1 7. (Original) A method as recited in Claim 5, wherein text similarity is
2 determined in part based upon weight values assigned to words of the text,
3 and wherein certain words have greater or lesser weight than other words.

1 8. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity include user click-through similarity.

1 9. (Original) A method as recited in Claim 8, wherein two documents are
2 considered similar based on user click-through similarity when the documents
3 are associated with similar patterns of user click behavior, selected from
4 among frequency of clicks, click context, duration of viewing, proximity in
5 time to other clicks, or proximity in context to other clicks.

1 10. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity are derived from patterns detected in user viewing of the
3 documents.

1 11. (Original) A method as recited in Claim 10, wherein the user viewing
2 information is monitored by a web caching system and stored in a log.

1 12. (Original) A method as recited in Claim 10, wherein two documents are
2 considered similar based on patterns of user viewing behavior, including
3 frequency of viewing, viewing context, duration of viewing, proximity in time
4 to other documents viewed by the same user, or similarity of patterns of
5 viewing by all users.

1 13. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity include URL similarity.

1 14. (Original) A method as recited in Claim 13, wherein two documents are
2 considered similar if a URL of each document contains similar URL sub-
3 components.

1 15. (Original) A method as recited in Claim 1, wherein the measures of document
2 similarity include multimedia similarity.

1 16. (Original) A method as recited in Claim 15, wherein two documents are
2 considered similar based on features derived from multimedia components
3 linked to or contained by the documents.

1 17. (Original) A method as recited in Claim 1, wherein the combination of two or
2 more measures of document similarity is achieved by taking the union of each
3 of a plurality of graphs, each graph describing one of the measures of
4 document similarity, to compute a combined graph that describes the
5 combined document similarity.

- 1 18. (Original) A method as recited in Claim 1, wherein the combination of two or
2 more measures of document similarity is achieved by taking the intersection
3 of each of a plurality of graphs, each graph describing one of the measures of
4 document similarity, to compute a combined graph that describes the
5 combined document similarity.
- 1 19. (Previously Amended) A method as recited in Claim 1, further comprising the
2 step of extracting similarity information from the similarity matrix to obtain
3 new documents supported by the set of training documents for each category.
- 1 20. (Previously Amended) A method as recited in Claim 19, wherein the
2 similarity information is obtained by optimizing an objective function.
- 1 21. (Previously Amended) A method as recited in Claim 19, wherein the
2 similarity information is obtained by only approximately optimizing an
3 objective function.
- 1 22. (Original) A method as recited in Claim 21, wherein approximately
2 optimizing the objective function comprises repeated application of a growth
3 transformation.
- 1 23. (Original) A method as recited in Claim 19, further comprising the step of
2 creating and storing a second matrix that represents an interim score for each
3 document in each category.

1 24. (Original) A method as recited in Claim 19, further comprising the steps of,
2 periodically as the matrix is being computed, normalizing rows of the matrix
3 by normalizing within each document, across all categories, whereby the score
4 for one document in a particular category will depend on the scores for that
5 document in all other categories.

1 25. (Original) A method as recited in Claim 19, further comprising the steps of,
2 periodically as the matrix is being computed, normalizing columns of the
3 matrix by normalizing within each category, across all documents, whereby
4 the score for one document in a particular category depends on the scores for
5 all other documents in that category.

1 26. (Original) A method as recited in Claim 1, in which the categories come from
2 a manually defined taxonomy.

1 27. (Original) A method as recited in Claim 1, wherein the categories are
2 derived from logs of user queries.

1 28. (Previously Amended) A method as recited in Claim 1, further comprising the
2 steps of creating and storing a second matrix using columns representing
3 documents and rows representing user sessions, and wherein values of
4 elements of the second matrix represent interest in a document shown by a
5 particular user in a particular session.

1 29. (Previously Amended) A method as recited in Claim 1, further comprising the
2 steps of creating and storing a matrix using columns representing user

3 sessions and rows representing documents, and wherein values of elements of
4 the second matrix represent interest in a document shown by a particular user
5 in a particular session.

1 30. (Original) A method as recited in Claim 28, wherein the element values are
2 computed as a function of a time that a user has spent viewing a document
3 associated with each element.

1 31. (Original) A method as recited in Claim 28, further comprising the steps of
2 creating and storing a second matrix representing a Similarity between pairs
3 of documents i and j, wherein the second matrix is derived by comparing pairs
4 of column vectors or row vectors, respectively i and j of the first matrix.

1 32. (Original) A method as recited in Claim 28, further comprising the steps of
2 creating and storing a second matrix representing a Similarity between pairs
3 of documents i and j, by finding pairs of documents i and j which have high
4 interest values for a particular user in a particular session or period of time.

1 33. (Original) The method recited in Claim 1, further comprising the steps of:
2 identifying a category of a classification taxonomy of the hypertext system in which a
3 first electronic document is presently classified; and
4 if a second electronic document is found to be highly Similar, storing information that
5 classifies the second electronic document into the category.

1 34. (Currently Amended) A computer-readable medium carrying one or more
2 sequences of instructions, wherein execution of the one or more sequences

3 of instructions by one or more processors causes the one or more
4 processors to perform the steps of:
5 establishing a plurality of training sets, wherein each training set is associated
6 with a category and includes training documents that have been classified
7 as belonging to said associated category;
8 determining how strongly each document of said plurality of documents
9 corresponds to each of said plurality of categories by determining
10 similarity between said each document and the documents that belong to
11 the training set of said category; and
12 wherein the step of determining similarity is performed using a matrix
13 representing document similarity that is derived by combining two or
14 more measures of document similarity and by one calculation of a
15 function referencing attributes associated with each of said plurality of
16 documents.

1 Claims 35 – 37 CANCELLED

2 38. (New) The method of claim 1, wherein the function is an objective
3 function and the one calculation is a maximization of the objective
4 function, where the objective function includes the matrix representing
5 document similarity.

1 39. (New) The method of claim 1, wherein the function referencing attributes
2 is performed by the function having terms with indices that are associated
3 with the plurality of documents.

1 40. (New) The method of claim 1, wherein the function referencing attributes
2 is performed by the function having confidence variables associated with
3 each of the plurality of documents, wherein the confidence variables
4 represent how strongly each of the plurality of documents corresponds to
5 the each of the plurality of categories.